The principles of the testing invariance measurement over time[1]

Kristýna Pospíšilová, Jiří Vinopal

Most social phenomena that are the subject of contemporary large national and international surveys are intended to be studied over the long term – the aim is to observe their development over time. A critical part of the observation of any social phenomenon over time is ensuring that the measurement of it remains invariant. However, this condition is not altogether easy to achieve in practice. Invariant measurement requires at the very least the following:

1. the standardisation of the research instrument and of the source of the data;

2. an instrument of sufficiently good quality (i.e. construct validity of the questionnaire);

3. a stable instrument, i.e. that it does not undergo (substantial) changes between individual measurements.

Prior to making comparisons of data from different years it is therefore always important to test the invariance of the measurements and thereby demonstrate that the given comparison will be valid. This is especially necessary if any changes have been made to the research instrument, such as changing the mode of data collection (a methodological change). Otherwise it is possible to expect a change in the understanding of key terms and the relations between them in the population itself (cultural and linguistic changes).

The procedure for testing an instrument's measurement invariance over time is analogical to that for testing invariance between groups (e.g. between countries in an international study).[2] In this case what we mean by groups are the individual occasions of data collection over time – for instance, the individual waves of a survey. Time series involve observing how a particular phenomenon develops over time within the same population, which is why there is a smaller likelihood of the invariance being disrupted as a result of cultural or linguistic differences, which otherwise tend to be the most common problem in international studies. But even these aspects cannot be ignored in long-term observations, as changes can occur in how terms and concepts are understood or a given society may ethnically transform, and so forth. Nevertheless, the bigger risk that we can expect in this case is the result of changes in the way the data collection is conducted, which naturally occur as part of continuous efforts to improve questionnaires and also occur as a result of changes in the methodology of the surveys, such as the gradual transition from PAPI to CAPI and CATI and eventually to CAWI in recent decades. This tutorial is based on an analysis of the invariance of an instrument, where the questionnaire underwent various changes over the course of ten years. The recommendations thus relate primarily to this area of possible changes, but the methods suggested below can also be applied analogically to other areas (e.g. testing invariance after a change in the mode of data collection).

Three main levels are discussed most often when considering measurement invariance. The lowest degree of comparability is represented by configural invariance (or construct invariance). In this case the data in every

[1] This work was supported by European Structural and Investments Funds, Operational Programme Research, Development and Education, project reg. no. CZ.02.1.01/0.0/0.0/16_013/0001796.
The text of this tutorial was adapted from the manuscript of the article *Measurement invariance of the instrument SQWLi over time*, submitted to the *Czech Sociological Review* in January 2020.

[2] Testing measurement invariance in cross-national studies in more detail for example: [Vandenberg and Lance 2000; Steinmetz 2013; Anýžová 2016; Flake and McCoach 2017]

group refer to the same social phenomenon – the same construct. If only this level of invariance is achieved, we know that the given set of items are indeed measuring the same phenomenon across every group, but we cannot make inter-group comparisons of the average item scores or even the relations between manifest or latent variables and other variables [Anýžová 2015].

A higher level is represented by metric invariance. The data attain this level when the scale range and the unit of measurement are identical, but the respondents in different groups perceive the scale range and unit in different ways. When metric equivalence is achieved it is still impossible to make inter-group comparisons of average item scores, but the relations of manifest or latent variables can be compared with other variables for which metric invariance was also attained [Anýžová 2015].

The highest level of measurement invariance is represented by scalar invariance. This is attained when the measurement scales used have the same range and the same unit and different groups of respondents also interpret the individual points on the scale in the same way. In this case it is possible to compare across groups the average item scores and the indexes they form. Given that the criteria for achieving scalar invariance are rather strict and often are not met, concepts have been developed that do not require perfect invariance – these are: partial scalar invariance and approximate measurement invariance.

In order for it to be possible to compare the indexes of individual items, domains, or two dimensions of the SQWLi over time, that is, to observe their development in time series, our data must demonstrate full scalar invariance, or at the very least partial scalar invariance or approximate measurement invariance. If they do not, there is a risk that the results for individual points in time will be distorted by excessive measurement error and the comparisons will be invalid.

The Instrument: SQWLi questionnaire

The SQWLi questionnaire (Subjective Quality of Working Life indicator) has been developed for the purpose of measuring trends in working-life quality in the Czech Republic over the long term. This means that it is to be used (primarily) to produce time series of an index for working-life quality and individual indexes for the different domains and dimensions of working life.[3]

The questionnaire has three sections in which respondents: (1) rate the importance of 18 aspects for their working life; (2) evaluate the same 18 aspects of their own working life; and (3) answer identification questions that can be used to analyse subgroups of workers. The SQWLi's basic indexes are computed from the first two sections (importance, evaluation), and separately in each dimension also for six domains and 18 aspects. All these indexes are to be monitored in the time series. Public access to the results is provided through a web application http://kvalitapracovnihozivota.vubp.cz/, for which all the indexes have been converted to a scale of 0 to 100.

In 2009 the content and structure of the instrument were already relatively firmly determined; nevertheless, even after 2009 numerous changes were made to the instrument, and their effects on measurement invariance may be quite diverse.

[3] The conceptualisation and development of the instrument has already been discussed and explained in greater depth in Vinopal [2011, 2012] and Vinopal and Čadová [2019].

Table 1. The structure of the SQWLi instrument

Domains	Aspects	Domains	Aspects
Reward	Level of earnings, pay	Time	How time-demanding the work is overall
	Fair reward		Distribution of working hours
	Earnings stability		Work doesn't interfere with personal time
Self-fulfilment	How interesting the work is	Conditions	Level of occupational health and safety
	Further education, personal development opportunities		Technical equipment used at work
	Job autonomy/independence		Workplace cleanliness, order, and hygiene
Relationships	Relationships between co-workers	Security	Nature of the employment relationship
	Superiors' behaviour towards subordinates		Job security
	Subordinates' behaviour towards superiors		Security in terms of employability

** Exact item wording of the Questionnaire is in Appendix*

The data and methods

In the analysis we include surveys from the year 2009 to 2019, a total of eight measurements. All the surveys were conducted by the Centre for Public Opinion Research at the Institute of Sociology, Czech Academy of Sciences, through its own interviewer network. The surveys were conducted using the PAPI method on representative quota samples based on current data from the Czech Statistical Office. The specific parameters of the survey in each individual year are presented in the table below.

Table 2. Surveys in the analysis

Year	Title	Data collection dates	Representativeness of the sample	N
2009	Stress in the Workplace…	22. 6. – 6. 7. 2009	Employees in the CR aged 18 to 65	836
2011	Czech Society 1102	7. 2. – 14. 2. 2011	Population of the CR over 15 years	563
2013	Czech Society 1306	3. 6. – 10. 6. 2013	Population of the ČR over 15 years	560
2014	Working-Life Quality 2014	19. 5. – 2. 6. 2014	Econ. active pop. CR aged 18+	2 029
2016	Quality of Life	31. 10. – 14. 11. 2016	Econ. active pop. CR aged 18+	750
2017	CSDA Research	18. 9. – 12. 10. 2017	Econ. active pop. CR aged 18+	675
2018	SQWLi optimisation–1st wave	26. 5. – 13. 6. 2018	Econ. active pop. CR aged 18+	1 018
2019	SQWLi optimisation-2nd wave	13.4. – 29.4.2019	Econ. active pop. CR aged 18+	478

Because the representativeness of all the samples was very good, no weights had to be added to the data, and given the number of cases we did not even proceed to imputation. The data from all the original scales were converted to the 0-100 scale.

Note: For the purpose of simplicity, only the first section – the battery of importance – is presented in this tutorial.

Two primary methods used to test measurement invariance nowadays are structural equation modelling (SEM), which is introduced in this tutorial, and item response theory (IRT), which is not presented here. In SEM the most frequently used approach is to test a set of increasingly restrictive models of multiple-group confirmatory factor analysis (MG CFA) and evaluate changes in the model fit statistics. First, a configural model is tested, to which the results of a metric model are then compared, and if they hold up, the next step is to test a scalar model (or then a partly scalar model) [Anýžová 2015].

Analysis

All the datasets must be first examined for representativeness and missing values. When there are problems with representativeness, weighting procedures might be introduced, and if there are larger shares of missing values the imputation might take place. If there were different scales used in individual years but but the differences between distributions are not so great, it may help to harmonise the scales on one range, for example 0 - 100.

The computations may be conducted in standard statistical programmes like SPSS, Mplus, R, or Stata (caution: not every programme is able to conduct all analyses presented here; ours were processed in IBM SPSS Statistics version 24 a Mplus 8.2).

Testing the conditions for measurement invariance

In all datasets the standard conditions for the use of EFA, CFA and SEM must be then examined: normal distribution, skewness and kurtosis, Kaiser-Meyer-Olkin measure (KMO)[4] Bartlett's Test for sphericity[5] and internal consistency (e.g. Cronbach's alpha). For determining comparability over time, however, the fundamental piece of information is the variability of *Cronbach's alpha if item deleted*, which should be as low as possible, indicating the stability of values over time and that none of the items stands out with any extreme values. [Anýžová 2015: 66]. The results might then look like those in the following table.

Table 3. Cronbach's Alpha analysis and KMO and Bartlett's Test (battery: importance)

	2009	2011	2013	2014	2016	2017	2018	2019	09-19	09-19
Cronbach's Alpha	.872	.898	.864	.878	.900	.889	.904	.916		
									min-max	var.
Cronbach's Alpha if item deleted									.853-.914	.051-.058
KMO	.858	.883	.833	.871	.897	.884	.908	.890		
Bartlett's Test	.000	.000	.000	.000	.000	.000	.000	.000		

In the next step, exploratory factor analysis must examine whether the data have a similar factor structure and the factor loadings are strong enough in every year. Of course, the factor structure must also be in concordance with the theory of the instrument or with other theoretical expectations. This analysis can

[4] KMO values ranges between 0 and 1. The closer the output is to 1 the better are the data. An acceptable minimum is considered to be a value of 0.6, and a value above 0.8 indicates very good data [Thompson 2004].
[5] This test must be statistically significant.

indicate problematic items or years: factor loadings below 0.3, large variation of factor loadings of individual items, or ambiguous affiliation of items with factors. The results might then look like those in the following table.

Table 4. EFA factor loadings (battery: importance)

		2009	2011	2013	2014	2016	2017	2018	2019	Var.
	A (earnings)	.775	.832	1.010	,826	,915	,858	,945	,861	,235
F1	B (fair reward)	.621	.672	.456	,569	,385	,672	,644	,650	,287
	C (ben./e. stab.)	.295	.202	.285	,289	,413	,284	,521	,492	,319
	D (co-workers)	.738	.716	.788	,722	,750	,877	,843	,810	,160
F2	E (superiors)	.822	.909	.850	,802	,754	,724	,811	,899	,186
	F (bull./sub.)	.519	.202	.581	,838	,832	,850	,937	,635	,735
	G (time demands)	.797	.736	.656	,803	,763	,771	,771	,893	,237
F3	H (time flex.)	.883	.947	.986	,851	,889	,862	,899	,873	,136
	I (harmonisation)	.554	.540	.666	.603	.416	.680	.704	.735	.319
	J (interestingness)	.629	.596	.511	.645	.721	.625	.705	.609	.210
F4	K (development)	.825	.920	.779	.882	.818	.780	.830	.892	.141
	L (independence)	.545	.643	.606	.679	.693	.702	.698	.630	.156
	M (contract)	.835	.599	.571	.638	.485	.618	.528	.660	.351
F5	N (security)	.614	.612	.730	.733	.689	.888	.599	.782	.289
	O (chances)	.168	.276	.492	.153	.348	.485	.528	.553	.399
	P (h&s)	.676	.580	.632	.745	.791	.757	.749	.793	.213
F6	Q (equipment)	.568	.850	.669	.676	.664	.844	.414	.494	.436
	R (hygiene)	.647	.634	.734	.667	.673	.671	.905	.766	.271

If in every year the analysis reveals the same factor structure that satisfactorily corresponds to the theory we can proceed to confirmatory factor analysis (CFA), which tests the original theoretical model or the model adjusted following the results of EFA (for example, by adding covariance between pairs of items). We can use the following standard fit statistics to assess the quality of the models: Chi-squared statistics enhanced by the number of degrees of freedom (CMIN/df), a comparative model fit index (CFI), the root mean square error of approximation (RMSEA), the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the standardised root mean square residual (SRMR).[6] We should give appropriate consideration to the fact that these statistics are often sensitive to at least one of the parameters of a given analysis – for example the complexity of the model or the size of the datasets and the number of groups in the comparison. The results might then look like those in the following table.

[6] A Chi-square test enhanced with the number of degrees of freedom should roughly have a value of 3. We can consider a model to be of adequate quality if its CFI index has a value of at least 0.9 (ideally 0.95), its RMSEA is up to 0.08 (it is sensitive to the complexity of the model), and the SRMR is up to 0.08 (ideally up to 0.06)) [Hu and Bentler 1999; van de Schoot et al. 2012; Byrne 2010].

Table 5. CFA models (battery: importance)

Year	CMIN	df	p	CMIN /df	CFI	RMSEA	RMSEA 90% CI	SRMR	sample size
2009	535	119	.000	4.496	.905	.070	.064 - .076	.061	716
2011	462	119	.000	3.882	.911	.078	.070 - .085	.066	475
2013	470	119	.000	3.950	.894	.076	.069 - .083	.059	509
2014	1267	119	.000	10.647	.914	.072	.069 - .076	.059	1838
2016	590	119	.000	4.958	.919	.075	.069 - .081	.054	702
2017	483	119	.000	4.059	.936	.070	.063 - .076	.054	633
2018	629	119	.000	5.286	.949	.067	.062 - .072	.042	949
2019	472	119	.000	3.966	.929	.080	.073 - .088	.049	453

In the example of the battery of importance, the tested model represents the real data well: the Chi-square test is significant, the CFI is greater than 0.9 in all the years except 2013 (0.894), the RMSEA is in every year equal to or less than 0.08, and the SRMR is also less than 0.08 and in fact never even rises above a value of 0.066. These results point to possible minor problems in data comparability (e.g. for the year 2013), but overall the model represents the data well in every year.

Invariance measurement: Multi-group confirmatory factor analysis (MG CFA)

In MG CFA a configural model is tested first, to which the results of a metric model are then compared, and if they hold up, the next step is to test a scalar model (or then a partly scalar model) [Anýžová 2015].

The MG CFA models used to test measurement invariance are usually assessed using the same model fit statistics as the CFA models. However, in the metric and scalar models, which are increasingly restrictive, we also turn our attention to the size of the change they exhibit (the change should not be greater than 0.01 in the CFI, 0.015 in the RMSEA, and 0.03 in the SRMR [Chen 2007; Hu and Bentler 1999; Anýžová 2015]). We also observe the AIC and BIC information criteria, which serve for a comparison of the models and which express the ratio between model fit and model complexity.[7] The results might then look like those in the following table.

Table 6. MG CFA models over time (battery: importance)

Model	CMIN	df	P	CMIN /df	CFI	RMSEA	RMSEA 90% CI	AIC	BIC	SRMR
Configural	4463	824	.000	5.416	.923	.075	(.073 - .077)	893952	897567	.056
Metric	4762	901	.000	5.285	.919 (.004)	.074 (.001)	(.072 - .076)	894098 (143)	897193 (-374)	.066
Scalar	7163	978	.000	7.324	.870 (.049)	.090 (.016)	(.088 - .092)	946999 (1010)	898921 (1728)	.099

[7] A lower IC value indicates a better model. If the change is greater than 10 points, this points to a significant worsening of the model [Anýžová 2015, van de Schoot et al. 2012].

Configural invariance refers to a situation where the number of latent variables and the structure of the factor loadings does not differ across groups. In the previous steps we tested that a given model with a specific number of factors presents the data from all the compared groups in a proportionate way and that the factor loadings are high enough (EFA) [Vandenberg and Lance 2000]. Then we tested the same model for each group separately (CFA). Now the model must be tested for all the groups together (MG CFA). In this baseline configural invariance model the same factorial pattern is specified for all groups with no other restrictions for loadings or intercepts. This model serves as the reference model, and it is with its model fit statistics that the other, more restrictive models are then compared [Meredith 1993; Byrne 2008].

In our case, the baseline configural model shows good results: the CFI is more than 0.9, the RMSEA is less than 0.08, and the SRMR is less than 0.06. Thus, the model with the given factor structure adequately represents the data in every measurement, despite some small imperfections in the model in individual years. This means that the instrument is indeed measuring the same construct across the years; and configural invariance is confirmed.

In the metric invariance model there is an additional requirement, the loadings were constrained to be equal across all groups being compared. The results of the metric model are compared with the configural model and if according to the changes in the model fit statistics they hold, it is possible to continue with further testing. In our metric model, the CFI is also greater than 0.9 and the change in the indicator is very low (0.004 is less than 0.01). The RMSEA is still acceptable (below 0.08) and its change is smaller than 0.015. The AIC rose significantly, but the BIC indicates an acceptable metric model, as does the SRMR, which is below 0.08 and the change is smaller than 0.03. The metric model thus still presents the data from all the compared groups well and metric invariance is also confirmed.

In the scalar model there is the additional requirement that the intercepts/thresholds of all the items were constrained to be equal across all groups being compared [Meredith 1993; Vandenberg and Lance 2000]. The results are compared to the metric model, and if they hold, the data can be used to compare groups. If they do not, it is possible to test partial scalar invariance which means searching for the most of the non-invariant items using modification indices (MI) and, where warranted, gradually releasing constraints on one or more loadings or intercepts or both for these item/s. We see a different situation here: all the model fit indicators in unison show that the scalar model is not acceptable. The CFI is below the minimum value of 0.9 and its change is greater than 0.1, the RMSEA is greater than 0.8, as is the SRMR. Full scalar invariance is thus not achieved.

In order to proceed further by testing partial scalar invariance we would have to gradually release, step by step and across all the groups, the restrictions of parameter correspondence for the items with the highest modification indices. This procedure must be repeated up until a satisfactory model is achieved but at the same time the majority of items on the factor should be invariant [Vandenberg and Lance 2000; Steinmetz 2013]. In cases like ours, this can be a lengthy process, and as Asparouhov and Muthén [2014: 495] have pointed out, these modifications can lead to the risk of producing an inappropriate model because of “the scalar model being far from the true model” and this procedure offers no guarantee that the simplest model and one easy to interpret will be achieved. Due to multicollinearity in the modification indices, the selection of the parameters to be freed is not unambiguous and thus other potentially better models can be overlooked. Another solution is to exclude the most problematic groups, but that has the effect of limiting further analytical options [Lomazzi 2018].

Seeking to resolve these limitations, Muthén and Asparouhov [2012, 2013] introduced the concept of ‘approximate measurement invariance’. While the procedure described above is premised on an expectation

of exact invariance of parameters, their concept rests on the assumption that some degree of non-invariance between parameters is acceptable and still allows making meaningful comparisons between groups. Their alignment method (AM) employs a simplicity function, which the authors liken to rotation in exploratory factor analysis. Using this method, it is possible to estimate all the model parameters in such a way that the number of non-invariant items and the size of the non-invariance are minimal.

Instead of partial scalar invariance, which in our case offers no promise of finding the simplest and a still well presentable model, we can decide to continue by testing approximate measurement invariance with the alignment method, which is recommended in cases like this by Muthén and Asparouhov [2014].

Invariance measurement: Alignment method

As its authors note, the main objective of the Alignment Method (AM) is to enable a comparison of factor means and variances without the need to achieve exact measurement invariance. The method does not require neither metric nor scalar invariance to be achieved – it is based on configural model. According to the authors it essentially automates and greatly simplifies measurement invariance analysis. Parameters are estimated so that they are comparable and the level of non-invariance is thus minimised [Muthén and Asparouhov 2014]. However, the authors recommend first testing for measurement invariance using traditional methods (configural, metric, and scalar models) and then comparing the results. The AM estimates the factor loadings and the intercepts of items for individual years. The results are presented in a table where the years in which the parameters are non-invariant are highlighted. The great advantage of AM thus lies in a detailed overview of the items which are the most invariant and which are most non-invariant over time. In order to obtain trustworthy alignment results and present meaningful comparisons between groups the non-invariant rate must not exceed 25%. The results might then look like those in the following table.

Table 7. Alignment method (battery: importance)

	Intercepts/Thresholds								Loadings							
	2009	2011	2013	2014	2016	2017	2018	2019	2009	2011	2013	2014	2016	2017	2018	2019
A (earnings)	ok	ok	ok	ok	X	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
B (fair reward)	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
C (ben./e. stab.)	ok	ok	ok	ok	ok	ok	X	X	ok	ok	ok	ok	ok	ok	ok	ok
D (co-workers)	ok	ok	X	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	X	ok	ok
E (superiors)	ok	ok	X	ok	ok	ok	ok	X	ok	ok	ok	ok	ok	X	ok	ok
F (bull./sub.)	ok	ok	X	ok	ok	ok	ok	X	ok	ok	ok	X	ok	ok	X	ok
G (time demands)	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
H (time flex.)	ok	ok	ok	X	X	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
I (harmonization)	X	ok	ok	X	ok	ok	ok	ok	ok	ok	ok	ok	ok	X	X	ok
J (interestingness)	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
K (development)	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
L (independence)	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
M (contract)	ok	ok	ok	ok	X	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok
N (security)	ok	X	X	X	ok	ok	ok	ok	ok	ok	ok	ok	ok	X	X	X
O (chances)	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	X	ok	ok	ok	ok

P (h&s)	ok	ok	ok	ok	ok	ok	ok	ok		ok	ok	ok	ok	ok	ok	ok	ok	
Q (equipment)	ok	ok	X	ok	ok	ok	ok	ok		ok	ok	ok	ok	ok	ok	X	ok	
R (hygiene)	ok	ok	ok	ok	ok	ok	ok	ok		ok	ok	ok	ok	ok	ok	ok	ok	

** Cases with non-invariant parameters are highlighted in grey.*

Out of the 288 parameters in total, 17 intercepts and 11 factor loadings in the importance battery are non-invariant, averaging to 9.7% non-invariance, well within the 25% cut-point. This is a very good result and it indicates that it is possible to compare latent means derived from alignment results between years [Muthén and Asparouhov 2014].

Results and conclusions

The analyses showed that the risk of invariance only begins to affect the SQWLi instrument's data on the scalar level, but the results of the AM tell us that even there it is not fatal. In any case, the given problems must be explained and the solutions to them enabling the instrument to be used in practice must be found on the basis of the individual items.

Table 8. Measurement invariance testing: overview (battery: importance)

	EFA	MI	AM
A (earnings)			
B (fair reward)			
C (ben./e. stab.)	X	X	
D (co-workers)			
E (superiors)			
F (bull./sub.)	X	X	X
G (time demands)			
H (time flex.)			
I (harmonisation)			X
J (interestingness)			
K (development)			
L (independence)			
M (contract)	X		
N (security)		X	X
O (chances)	X	X	
P (h&s)			
Q (equipment)	X		
R (hygiene)			

** The table does not contain a record of all the problems that were detected and shows only those problems that proved to be the biggest ones within the frame of this analysis.*

It might happen that in the course of time an item has changed considerably, including a change in the concept it measures. Here this is the case of two items (C and F). Although the different versions of these items always work well enough in the overall construct, after a change in concept they in themselves are then measuring something different. Even without an analysis of invariance it is clear that it makes no sense to observe the

items separately over time as a single phenomenon. Therefore, even when the total results of an invariance analysis are good, that should not deter us from thoroughly validating them by closely controlling for the specific nature of the items in all the groups. If, for example, we were working with an instrument that we were not familiar with in detail or did not have detailed documentation on, we could very easily commit fundamental errors.

Nevertheless, when we confirm problems with the items in the invariance models, this does not automatically call the instrument as a whole into question. The fact that even when substantive changes are made to two items the construct as a whole and its individual factors still function in a satisfactory manner supports the conclusion that for the observed years it is possible to treat the instrument as invariant.

For making comparisons in time we then have more options. We can exclude problematic item(s) or problematic year(s) form the comparison. Another solution is not to compare the scores of single items, but only to compare the composite scores of the whole factors proved to be invariant (e.g. factor scores). Which option we choose depends mainly on the goal of a given comparison and on how much data (items, years) we have.

References

Anýžová, P. 2015. *Srovnatelnost postojových škál v komparativním výzkumu.* Olomouc: Univerzita Palackého v Olomouci.

Asparouhov, T. and B. Muthén. 2014. 'Multiple-Group Factor Analysis Alignment.' *Structural Equation Modeling: A Multidisciplinary Journal* 21: 495-508, http://dx.doi.org/10.1080/10705511.2014.919210.

Byrne, B. M. 2010. *Multivariate applications series. Structural equation modeling with AMOS: Basic concepts, applications, and programming (2nd ed.).* Routledge/Taylor & Francis Group.

Chen, F. F. 2007. 'Sensitivity of goodness of fit indexes to lack of measurement invariance.' *Structural Equation Modeling: A Multidisciplinary Journal* 14 (3): 464–504, https://doi.org/10.1080/10705510701301834.

Flake J. K. and D. B. McCoach. 2017. 'An Investigation of the Alignment Method With Polytomous Indicators Under Conditions of Partial Measurement Invariance.' *Structural Equation Modeling: A Multidisciplinary Journal* 25 (1): 56–70, https://doi.org/10.1080/10705511.2017.1374187.

Hu, L.-T. and P. M. Bentler. 1999. 'Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives.' *Structural Equation Modeling: A Multidisciplinary Journal* 6 (1): 1–55, https://doi.org/10.1080/10705519909540118.

Lomazzi, V. 2018. 'Measurement Invariance of Gender Role Attitudes in 59 Countries.' *Methods, data, analyses* 12 (1): 77-104, http://dx.doi.org/10.12758/mda.2017.09.

Meredith, W. 1993. 'Measurement invariance, factor analysis and faktorial invariance.' *Psychometrika* 58: 525–543, https://doi.org/10.1007/BF02294825.

Muthén, B., T. Asparouhov. 2012. 'Bayesian structural equation modeling: A more flexible representation of substantive theory.' *Psychological Methods* 17: 313-335, https://doi.org/10.1037/a0026802.

Muthén, B., T. Asparouhov. 2013. '*New Methods for the Study of Measurement Invariance with Many Groups.*' Technical report. Retrieved 14 January 2020 (http://www.statmodel.com/download/PolAn.pdf).

Muthén, B., T. Asparouhov. 2014. 'IRT studies of many groups: the alignment method.' *Frontiers in psychology* 5: 978, https://doi.org/10.3389/fpsyg.2014.00978.

Steinmetz, H. 2013. 'Analyzing Observed Composite Differences Across Groups Is Partial Measurement Invariance Enough?' *Methodology* 9 (1): 1-12. https://doi.org/10.1027/1614-2241/a000049.

Thompson, B. 2004. Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications. Washington: American Psychological Association. https://doi.org/10.1037/10694-000 .

van de Schoot, R., P. Lugtig, J. Hox. 2012. 'A checklist for testing measurement invariance.' *European Journal of Developmental Psychology* 9 (4): 486-492, https://doi.org/10.1080/17405629.2012.686740.

Vandenberg, R. J., C. E. Lance. 2000. 'A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research.' *Organizational Research Methods* 3: 4-70. https://doi.org/10.1177/109442810031002.

Vinopal, J. 2011. 'Indikátor subjektivní kvality pracovního života.' (Subjective quality of working life indicator) *Sociologický časopis/Czech Sociological Review* 47 (5): 937 - 965.

Vinopal, J. 2012. 'The Discussion of Subjective Quality of Working Life Indicators.' *Sociológia* 44 (3): 385-401.

Vinopal, J. and N. Čadová. 2019. 'Povaha aspektů pro měření objektivní kvality pracovního života.' (Nature of aspects for measuring objective quality of Working life) *Časopis výzkumu a aplikací v profesionální bezpečnosti* 12 (spec.: Nové trendy v BOZP 2019). Retrieved 14 January 2020 (https://www.bozpinfo.cz/josra/povaha-aspektu-pro-mereni-objektivni-kvality-pracovniho-zivota. ISSN 1803-3687).

Zercher F., P. Schmidt, J. Cieciuch, E. Davidov. 2015. 'The comparability of the universalism value over time and across countries in the European Social Survey: exact vs. approximate measurement invariance.' *Frontiers in psychology* 6: 733, https://doi.org/10.3389/fpsyg.2015.00733.

Appendix: Subjective Quality of Working Life indicator (SQWLi), Version SQWLi_19b_P1 (May 2019)

"Imagine, please, that you are currently deciding on a new job. For every aspect I am going to read to you, tell me how important or unimportant it is for you personally. Use a range from 0 to 10, where 0 stands for FULLY UNIMPORTANT and 10 for FULLY ESSENTIAL.

FULLY UNIMPORTANT										FULLY ESSENTIAL	DO NOT KNOW
0	1	2	3	4	5	6	7	8	9	10	99

h) The distribution of working hours during the day or week.

g) Total duration of working hours.

i) So that your work does not interfere with your personal time, i.e. time for family, interests or relax.

n) To be sure you don't lose your job.

o) So that your work gives yozu the chance of further possible employment in the labour market.

m) The nature of the employment relationship, i.e. whether you have a permanent or fixed-term contract, a full-time or part-time contract, whether you work as an employee or a self-employed, etc.

c) Earnings stability, so that your salary is regular and stable.

a) The amount of earnings, i.e. the amount of your salary or wages.

b) So that your work results are financially rewarded in a fair way.

p) The level of occupational health and safety.

r) Cleanliness, tidiness and hygiene at work.

q) Technical equipment for work.

d) Relationships between co-workers.

e) Behaviour of superiors towards subordinates.

e) Behaviour of subordinates towards superiors.

l) To have the opportunity to decide on your own work tasks, organize your work independently.

k) To have opportunities for further education and personal development at work.

j) To have interesting work.

“Now I will again read aspects of working life, as before. But this time, evaluate whether your current main job is bad or good in that respect. Use the range from -5 to +5, where -5 means VERY BAD

VERY BAD										VERY GOOD	NOT APPLICABLE	DO NOT KNOW
-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	88	99

h) The distribution of working hours of the main job during the day or week.

g) Total duration of working hours of your main job.

i) How your work does interfere with your personal time, i.e. time for family, interests or relax.

n) How sure you are that you don't lose your job.

o) What chance of further possible employment in the labour market does your work give you.

m) The nature of your employment relationship, i.e. whether you have a permanent or fixed-term contract, a full-time or part-time contract, whether you work as an employee or a private person, etc.

c) Earnings stability, i.e. how regular and stable your salary is.

a) The amount of earnings, i.e. the amount of your salary or wages.

b) The fairness of the financial reward of your work results.

p) The level of occupational health and safety in your main job.

r) Cleanliness, tidiness and hygiene at work.

q) Technical equipment for your work.

d) Relationships between co-workers.

e) Behaviour of superiors towards subordinates.

e) Behaviour of subordinates towards superiors.

l) To have the opportunity to decide on your own work tasks, organize your work independently.

k) Which opportunities for further education and personal development does this job give you.

j) How interesting your main work is.