Measurement Invariance of the SQWLi Instrument Over Time*

JIŘÍ VINOPAL and KRISTÝNA POSPÍŠILOVÁ**
Institute of Sociology, Czech Academy of Sciences, Prague

Abstract: The SQWLi questionnaire was developed for the long-term measurement of subjectively perceived quality of working life. The aim of this study is to test the instrument's measurement invariance between 2009 and 2019 and determine whether – despite the modifications made to the instrument over the years – the results remain comparable. Data from eight representative surveys of the economically active population in the Czech Republic were analysed (total N = 6909) using the MG CFA method (configural, metric, and scalar invariance) and the alignment method (approximate measurement invariance). The findings from the MG CFA tests for measurement invariance indicate that the SQWLi instrument achieves configural and metric invariance over time but not full scalar invariance. Achieving a partial scalar invariance would be challenging because of the many high modification indices; therefore, an approximate measurement invariance approach, namely the Alignment Method, was applied. The results suggest that comparisons of latent means across all years can be made. Consequently, it is possible to make meaningful comparisons of overall indices of dimensions (batteries) and of more general domains. However, not all the individual items can be compared. The results confirm that the biggest risk of invariance is caused by conceptual changes to items and by substantial or frequent modifications to item wording. Conversely, the results also show that a conceptual change to an entire dimension may not necessarily cause any problem on the general level, and that a disruption of invariance caused by changes to the range of scales used can be rectified by means of their harmonisation ex post.

Keywords: SQWLi, subjective quality of working life, exact and approximate measurement invariance, alignment method

Sociologický časopis/Czech Sociological Review, 2021, Vol. 57, No. 3: 343–375
https://doi.org/10.13060/csr.2020.048

* Acknowledgements: 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.

** Direct all correspondence to: Jiří Vinopal, Institute of Sociology, Czech Academy of Sciences, Jilská 1, 110 00 Prague 1, e-mail: Jiri.Vinopal@soc.cas.cz; Kristýna Pospíšilová, Institute of Sociology, Czech Academy of Sciences, Jilská 1, 110 00 Prague 1, e-mail: kristyna.pospisilova@soc.cas.cz.

The SQWLi questionnaire was developed for the long-term measurement of subjectively perceived quality of working life. It is to be used (primarily) to produce time series of various indices of working-life quality and its partial dimensions or domains.

A critical part of the observation of any social phenomenon over time is equivalent method. This study seeks to examine whether measurement invariance and consequently also the comparability of data in a time series are adversely affected when changes are made to the instrument of measurement; and if so, how large an effect the various modifications have on measurement invariance. The analyses presented in this paper (1) provide information on the quality of the SQWLi instrument itself and how suitable its data are for creating time series and (2) on a more general level provide an idea of what effect any modifications to the research instrument may have on the quality of the time series. They will also show that invariance can be increased ex post by harmonising the measurement scales.

Measuring the (subjective) quality of working life

For decades now there have been efforts to observe working-life quality in various contexts and on different levels. At the country level, the quality of working life is one dimension that figures in political debates about the direction of the economy, the labour market and labour regulations, etc. In this area it is one of the macro indicators cited by various parties to policy negotiations, and in this function it must necessarily assume the fullest, most aggregate form possible. At the opposite end of the spectrum, working-life quality is addressed as an element in the employee policies of individual employers, where it figures as one aspect of employee care – for example, in efforts to minimise fluctuation, improve work performance, or maintain social harmony. In such cases it is not about describing or analysing large populations, but rather about closely analysing individual workers for targeted, practical effects. In between these macro and micro levels, there are various motives and methods behind efforts to survey the quality of working life or its alternative or related conceptualisations such as job satisfaction, job stress, working conditions, decent work, etc. [Hoppock 1935; Cranny et al. 1992; Danna and Griffin 1999; Sirgy et al. 2001; Ghai 2008; Judge et al. 2012].

From this perspective, the SQWLi is an instrument of the macro level, where measuring working-life quality has its own long tradition [e.g. Sirgy et al. 2001; Tangian 2005; Lowe 2007; Leschke et al. 2008; Fuchs 2009; Swamy et al. 2015] and is currently also being pursued through many different avenues. The biggest projects that deal directly with working-life quality or that serve as the most common source of data on the subject are the European Working Conditions Surveys (EWCS), European Union Statistics on Income and Living Conditions (EU-SILC), and the European Labour Force Survey (ELFS); but there is also, for example,

Eurobarometer and the Gallup Poll. Special attention should be paid to the Arbeitsklima index in Austria and the DGB-Index Gute Arbeit in Germany, which are projects that have a purpose, method, and underlying principles similar to SQWLi. The most systematic conceptualisation and overview of the instruments used to measure working-life quality is by de Bustillo et al. [2011]. In their analysis they show that many indices suffer from poorly conceptualised combinations of objective and subjective characteristics or (relatedly to some extent) from the mixing of macro-, mezzo-, and micro-level data. The authors also demonstrate that different indices capture different areas of working life, that they combine in varying proportions the characteristics of employment and the characteristics of the work itself, the procedures and results, and are aimed at producing either a summary index or a system of indicators. The SQWLi questionnaire is not intended to bridge all the varieties of instruments and possibilities presented here; our aim is to offer a good conceptualisation of working-life quality as a subjective concept and to tweak and fine tune an instrument that can be meaningfully used to monitor large populations over the long term.

A description of the SQWLi instrument

The SQWLi questionnaire was created in a gradual process of development and testing that has been under way since 2005. First, after analysing 66 aspects of working life that were examined in extensive research in 2005, 18 of these aspects of working life were selected and organised into 6 domains.[1] In subsequent years the instrument underwent methodological improvements to address specific problems and enhance overall quality.

The questionnaire has three sections in which respondents: (1) rate the *importance* of 18 aspects of working life for them; (2) *evaluate* the same 18 aspects of their own working life; and (3) answer background socio-demographic questions.[2] The SQWLi's basic indices (converted to a scale of 0 to 100) are computed for the two dimensions (*importance* and *evaluation*), and separately in each dimension also for six domains and 18 aspects. All these indices are to be monitored in time series.[3]

Our analysis of invariance includes only surveys conducted since 2009, by which time the content and structure of the instrument were already relatively firmly determined. Nevertheless, even after 2009 numerous changes were made with supposedly diverse effects on measurement invariance. Table 2 provides an

[1] For more on the development and the properties of the instrument, see Vinopal [2011].
[2] This conceptualisation has already been discussed and explained in greater depth in Vinopal [2011, 2012] and Vinopal and Čadová [2019].
[3] Publicly accessible through a web application: http://kvalitapracovnihozivota.vubp.cz/ (retrieved 14 January 2020).

Table 1. The structure of the SQWLi instrument (identical in the two dimensions: importance and evaluation)

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

overview of the types and extent of the modifications, which ranged from minor shifts in the wording of items or the question to changes in the range or polarity of the scales and changes in the theoretical concept of individual items or even an entire dimension. Once modifications were made to the instrument, they were retained and we only very rarely went back to any previous version. Therefore, the changes accumulated over years in total.

One of the most important conceptual changes was made between the first and second surveys, when the concept of the second battery switched from *satisfaction* (satisfied – dissatisfied) to *evaluation* (bad – good). Other fundamental changes were made in 2014: the rating scales were substantially modified (the polarity was reversed, the range was increased from 6 to 11 points, the unipolar scale for rating *importance* was made numerically distinct from the bipolar scale for the *evaluation*, and the labels were reduced to just the end points); the concept of item F (bullying/subordinates) was changed; and substantive or wording changes were made to most of the other items. In 2018 conceptual changes were made to item C (benefits/earnings stability) and in 2019 again to item F (bullying/subordinates). Not entirely minor modifications were also made in 2018

Table 2. An overview of modifications made to the instrument in individual years (changes marked in grey shading) – first part

Object of change	Type of change	2009	2011	2013	2014	2016	2017	2018	2019
IMPORTANCE	Concept	Importance	Importance	Importance	Importance	Importance	Importance	Importance	Importance
Question	Wording of the question	Short	Short	Short	Long	Long	Long	Long	Short
	Substantive change in the wording	X	YES	YES	YES	NO	NO	YES	NO
	Linguistic change in the wording	X	YES	YES	YES	NO	NO	YES	YES
Item (number of items changed)	Change in item concept	X	0	0	1[1]	0	0	1[2]	1[3]
	Substantive change in the wording	X	4	0	11	3	4	3	2
	Linguistic change in the wording	X	5	3	11	4	6	12	18
Scale	Scale range	6	6	6	11	11	11	11	11
	Scale values	1 6	1 6	1 6	0 10	0 10	0 10	0 10	0 10
	Scale labels	All points	All points	All points	End points	End points	End points	End points	End points
	Scale polarity	+/–	+/–	+/–	–/+	–/+	–/+	–/+	–/+
	Labels	Definitely important – Definitely unimportant	Definitely important – Definitely unimportant	Definitely important – Definitely unimportant	Completely unimportant – funda-mentally important	Completely unimportant – funda-mentally important	Completely unimportant – funda-mentally important	Completely unimportant – funda-mentally important	Completely unimportant – funda-mentally important

Table 2. An overview of modifications made to the instrument in individual years (changes marked in grey shading) – second part

Object of change	Type of change	2009	2011	2013	2014	2016	2017	2018	2019
IMPORTANCE	Concept	Importance	Importance	Importance	Importance	Importance	Importance	Importance	Importance
Administration	Order of the domains	09_P	09_P	09_P	09_P	09_P	09_P	18a_P4	18a_P4
	Order of the items	09_P	09_P	09_P	09_P	09_P	09_P	18a_P4	18a_P4
EVALUATION	Concept	Satisfaction	Evaluation	Evaluation	Evaluation	Evaluation	Evaluation	Evaluation	Evaluation
Question	Wording of the question	Short	Short	Short	Long	Long	Long	Long	Short
	Substantive change in the wording	X	YES	YES	YES	NO	NO	NO	NO
	Linguistic change in the wording	X	YES	YES	YES	NO	NO	YES	NO
Items (number of items changed)	Change in item concept	X	0	0	1[1]	0	0	1[2]	1[3]
	Substantive change in the wording	X	5	0	17	3	4	4	2
	Linguistic change in the wording	X	11	1	17	4	6	9	18

Table 2. An overview of modifications made to the instrument in individual years (changes marked in grey shading) – third part

Object of change	Type of change	2009	2011	2013	2014	2016	2017	2018	2019
IMPORTANCE	Concept	Importance	Importance	Importance	Importance	Importance	Importance	Importance	Importance
Scale	Scale range	6	6	6	11	11	11	11	11
	Scale values	1 6	1 6	1 6	–5 +5	–5 +5	–5 +5	–5 +5	–5 +5
	Scale labels	All points	All points	All points	End points	End points	End points	End points	End points
	Scale polarity	+/–	+/–	+/–	–/+	–/+	–/+	–/+	–/+
	Labels	Very satisfied – very dissatisfied	Very good – very bad	Very good – very bad	Very bad – very good	Very bad – very good	Very bad – very good	Very bad – very good	Very bad – very good
Administration	Order of the domains	09_P	09_P	09_P	09_P	09_P	09_P	18a_P4	18a_P4
	Order of the items	09_P	09_P	09_P	09_P	09_P	09_P	18a_P4	18a_P4

Note: The record of changes made to the questionnaire presented in the table refers in each case to the immediately previous version of the questionnaire; if no change is indicated, this means that there was no change from the preceding version.

[1] Change in the concept of item f): violence and bullying in the workplace – the overall nature of relations between co-workers at the workplace.

[2] Change in the concept of item c): nonfinancial advantages/bonuses/benefits – earnings stability, that earnings are regular and steady.

[3] Change in the concept of item f): the nature of relations between co-workers in the workplace – subordinates' behaviour towards superiors.

when the order of the domains and items were changed, and the wording of the question and items were reduced in length.

It is not possible in this paper to explain all the reasons, empirical evidence, or actual effects of the changes that were made. Generally, changes to the concepts or wording of items were in almost every case motivated by problems of an empirical nature (higher item nonresponse, weaker discrimination, weaker performance in the EFA models, weaker regression coefficients in the CFA, etc.). Changes to the scales were made as we advanced our expertise about the relationship between types of concepts, rating scales, and measurement quality [Saris and Gallhofer 2014]. The concept of *satisfaction* was substituted for *evaluation* in 2011 following discussions at international meetings with colleagues also focusing on the measurement of working-life quality and based on the theoretical assumption (to some extent also supported by empirical evidence [Vinopal 2009]) that when people are rating their satisfaction with something, they are already intuitively considering that something's importance at the same time; and thus the concept of *importance* would be essentially measured twice. Finally, the change of the order of the domains and items in 2018 was done after the experiment in which the instrument was prepared for the possibility of rotation in electronic data collection modes. The results showed that one of the new tested orderings produced statistically better models, so it was retained for the future to be used also in PAPI.

Measurement invariance

Measurement invariance (also known as equivalence of measurement) is a prerequisite for making meaningful comparisons of differences in a construct between groups when item-based scales are used [Flake and McCoach 2017]. We most often encounter the issue of measurement invariance in the context of international comparative studies, but it is just as important in comparisons over time [Davidov 2010].

In principle, invariance should be a concern whenever the aim is to make a comparison between groups measured with the same instrument. Most often such groups are countries or socio-demographic groups. But still more often measurement invariance is applied also in the case of repeated research. In such cases, groups are represented by points of measurement in time. Here the term 'groups' refers to samples from a population whose characteristics may have changed because of natural developments; i.e. groups are usually represented by waves of data collection (years in this article). In all such cases it is crucial to ensure that the respondents in every group (year in this case) understand and interpret the measurement instrument in the same way; if they do not, it is impossible to make meaningful comparisons between groups.

There exist multiple approaches for testing measurement invariance. In this article we focus on confirmatory factor analysis (CFA and MG CFA). This ap-

proach is based on the logic of hierarchical model comparisons, where model fit statistics for every model are assessed and compared to the ones of the following, more restrictive model [Meredith 1993]. Three main levels are discussed most often when considering measurement invariance: configural, metric, and scalar invariance.

The lowest degree of comparability is represented by configural invariance (or construct invariance). In this case the data in every group refer to the same social phenomenon – the same construct – even though they were not necessarily obtained using identically formulated questions or identical measurement scales with the same range. 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 invariance is achieved it is still impossible to make inter-group comparisons of average item scores, but it is possible to make meaningful comparisons of relations of latent variables with other manifest 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 indices 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, the latter of which will be discussed later.

So that we have some methodological support for making comparisons of the individual items' means, factors' means, domains, or two dimensions of the SQWLi over time, that is, for following 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 research question: expected problems with measurement invariance

The primary research question we are concerned with here is whether the SQWLi instrument retains measurement invariance over time even if modifications have been made to it over the course of the years. More specifically, we want to deter-

mine what level of invariance can be achieved across all the years compared (2009 to 2019), and between which years and which items the resulting indices of the SQWLi can be compared over time. Based on the changes made to the instrument described above, we expect to encounter the following problems on each of the individual levels of measurement invariance:

Configural invariance

The conceptual change of the whole dimension (battery) from *satisfaction* to *evaluation* in 2009 could in theory interfere with the configural invariance. However, the results of the continuous analyses conducted in the course of developing the instrument did not signal any significant changes in the factor structure. Consequently, we no longer expect this to be a problem.

In 2018 the order of the domains and the order of the items were changed, and numerous wording modifications were also made along with some substantive changes. However, even these changes did not have an effect on the shape of the factor structures of the two batteries, which were verified after each survey. We therefore do not expect these changes to disrupt the configural invariance either.

Metric invariance

A change made to the rating scales in 2014 (from a 6-point to an 11-point scale) had the potential to disrupt the metric invariance. As we were to some extent able to determine in a previous study [Šeflová 2016], this did indeed occur. In order to make comparisons, we would therefore have to exclude the data from 2009, 2011, and 2013, and that would mean losing five years of the time series conducted to date. This fact prompted us to search for a solution in our current study that would allow us to keep these years in our comparisons. Based on an analysis of the response distributions and total indices of items on both versions of the scales used, we therefore decided to convert all the data to the same scale, which had already been demonstrated elsewhere to be a functional solution [Šeflová 2016]. The scale that suggested itself was 0–100, used to present the indices in our time series on-line. With this harmonisation we expect metric invariance will be achieved across all the years.

Scalar invariance

Over the years at least one change was made to the wording of every item in the questionnaire; most of the items, however, had several changes made to them, and some were even more extensively altered. In our view, the most problematic

changes were those that involved a change in theoretical concept: F (bullying/subordinates)[4], C (benefits/earnings stability);[5] or more substantive or more frequent changes in item wording: E (superiors), L (independence), M (contract), N (security), O (chances). (The wording of all the items is given in Appendix 1.) Given the many changes made to the items, we expect that they will have an adverse effect on full scalar invariance, and that this will moreover be caused by various different items in different years. Instead of testing models of partial scalar invariance, we apply the approximate invariance approach and use the alignment method (which is recommended in such a situation). We expect approximate measurement invariance to be confirmed, or that it will be disrupted only by the items that have undergone a change in theoretical concept.

Data

In our 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 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.

Because of the lack of significant deviation of the parameters of samples from the parameters of the population, there was no need to weight the data. Given the number of cases, we did not even proceed to item-nonresponse imputations. The data from all the original scales were converted to the 0–100 scale.

Testing measurement invariance for quota samples is not common because of their non-probabilistic nature. In probabilistic samples, one can expect measurement invariance across samples just because the population is 'standardised' by the random sampling procedure. One objection might be that, in the case of quota samples, the selection of individuals depends much more on the researcher than the procedure. Here the interpretation of observing measurement invariance may be quite similar: in our surveys the procedure of (quota) sampling was still the same; therefore, if nothing else changed, it should produce invariant outputs. It could be said that the (quota) samples in this study are 'standardised' by the constant design of the sampling procedure used.

[4] Violence and bullying in the workplace / the overall nature of relations between coworkers in the workplace / subordinates' behaviour towards superiors.

[5] Nonfinancial advantages/bonuses/benefits / earnings stability, that earnings are regular and steady.

Table 3. 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 CR 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

Methods

One of the primary methods used to test measurement invariance is structural equation modelling (SEM), for which we are able to mathematically define various levels of measurement invariance [Hirschfeld and von Brachel 2014]. 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 partial scalar model) [Anýžová 2015].

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 first step we have to test that a given model with a specific number of factors presents the data from all the compared groups (years) in a proportionate way and that the factor loadings are high enough [Vandenberg and Lance 2000]. The model is tested first for each group separately and then for all the groups together. In this baseline configural invariance model the same factorial pattern was 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 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 the scalar model there is the additional requirement that the intercepts/thresholds of all the items are constrained to be equal across all the 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 most of the non-invariant items using modification indices (MI) and, where warranted, gradually releasing the constraints on one or more loadings or intercepts or both for these item/s. This procedure must be repeated 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].

Given that the likelihood of confirming full scalar invariance decreases with the increasing number of groups included in a comparison [Zercher et al. 2015], researchers are often forced to search for a partially scalar model. However, 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. Owing to the 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 us to make 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. As Muthén and Asparouhov [2014: 3] state: 'As a rough rule of thumb, a limit of 25% non-invariance may be safe for trustworthy alignment results.' When the share of non-invariant parameters is below this limit it is possible to compare factor means and factor variances across all groups.

Partial scalar invariance expects exact comparability for the majority of the parameters and allows non-invariance for a small number of them. The alignment method permits a small deviation for a larger number of parameters and therefore leads to better results in most situations [Zercher et al. 2015]. In our

study the SQWLi instrument is primarily tested using the MG CFA method. Given the number of modifications made to the items, which we do not, however, regard as particularly pronounced with respect to the consistency of the construct as a whole, instead of testing partial scalar invariance we applied the alignment method to test the instrument.[6]

Testing the conditions for measurement invariance

All the datasets were first examined for the share of missing values and a normal distribution. In all the years observed the share of missing values was around 4%, and although the data exhibit a certain level of skewness and kurtosis this is not an impediment to further analysis (over the years the interval of skewness ranged between –2.444 and –0.897 and the interval of kurtosis ranged between –0.543 and 8.573).

The first information about potential problems with the comparability of data over time can usually be obtained by analysing the scale's reliability, which can be tested using Cronbach's alpha. For both of our batteries the alpha is very high in all the observed years, which indicates that the scale is reliable. The alpha is also relatively stable over time and slightly increasing, which could indicate that the modifications made to the SQWLi were to its benefit. For determining comparability over time, however, the fundamental piece of information is the variability of Cronbach's alpha if item deleted, which is for all items close to zero (max. 0.058). The values are thus very stable over time, despite the changes made to the instrument, and none of the items stands out with any extreme values or indicates any major invariance problems. [Anýžová 2015: 66]

That the preconditions for using factor analysis were met was determined using the Kaiser-Meyer-Olkin measure (KMO) and Bartlett's Test of sphericity. The values of these indicators were also found to be good, as in every case the KMO is more than 0.83 and Bartlett's Test is always significant. All the datasets from all the observed years are thus suited to the use of factor analysis.

[6] The calculations were performed in IBM SPSS Statistics version 24 and Mplus 8.2. For the CFA and the alignment method we used the ML (maximum likelihood) estimate function.

Table 4. Cronbach's alpha analysis

	Cronbach's alpha								C alpha if ID	
	2009	2011	2013	2014	2016	2017	2018	2019	min – max	Var.
Importance	.872	.898	.864	.878	.900	.889	.904	.916	.853–.914	.051–.058
Evaluation	.921	.921	.911	.930	.909	.913	.922	.938	.901–.937	.028–.031

Table 5. KMO and Bartlett's Test

		2009	2011	2013	2014	2016	2017	2018	2019
Importance	KMO	.858	.883	.833	.871	.897	.884	.908	.890
	Bartlett's Test	.000	.000	.000	.000	.000	.000	.000	.000
Evaluation	KMO	.897	.896	.900	.916	.890	.893	.914	.900
	Bartlett's Test	.000	.000	.000	.000	.000	.000	.000	.000

Table 6. EFA factor loadings: Importance

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

Exploratory Factor Analysis (EFA)

We examined whether the data have in every year a similar factor structure and factor loadings that are strong enough. An exploratory factor analysis was applied to both batteries and in the first step for each year separately.[7]

In every case the analysis revealed a structure that corresponds satisfactorily with the theory; it also, however, pointed to items that we may expect to show problems with invariance. These could be items that have a factor loading below 0.3 and a variability of factor loadings that is greater than 0.3. For the *importance* battery this primarily concerns items C (benefits/earnings stability), F (bullying/subordinates), M (contract), O (chances), and Q (equipment); for the *evaluation*

[7] Principal Axis Factoring extraction method was used with Oblimin rotation method and Kaiser Normalisation. The number of factors was fixed to 6 according to the construct although the sixth factor usually has the eigenvalue below 1.0. Previously discussed in Šeflová [2016].

Table 7. EFA factor loadings: Evaluation

		2009	2011	2013	2014	2016	2017	2018	2019	Var.
F1	A (earnings)	.740	.578	.876	.927	.883	.868	.833	.970	.391
	B (fair reward)	.940	.630	.703	.853	.703	.931	.794	.671	.310
	C (ben./e. stab.)	.355	.593	.406	.496	.606	.501	.210	.217	.395
F2	D (co-workers)	.626	.646	.631	.775	.701	.774	.737	.878	.253
	E (superiors)	.555	.763	.727	.648	.661	.668	.667	.694	.208
	F (bull./sub.)	.839	.533	.579	.977	1.038	.922	.882	.797	.505
F3	G (time demands)	.920	.796	.803	.745	.782	.840	1.004	.970	.258
	H (time flex.)	.863	.825	.845	.837	.914	.834	.751	.843	.163
	I (harmonisation)	.785	.851	.746	.866	.836	.834	.786	.804	.120
F4	J (interestingness)	.828	.824	.657	.658	.774	.757	.659	.457	.371
	K (development)	.691	.644	.783	.953	.788	.868	.868	.775	.309
	L (independence)	.477	.597	.515	.635	.688	.574	.581	.566	.211
F5	M (contract)	.186	.491	.602	.582	.455	.552	.578	.914	.728
	N (security)	.969	.768	.860	.928	.909	.924	.739	.509	.459
	O (chances)	.575	.315	.392	.316	.262	.449	.651	.386	.389
F6	P (h&s)	.752	.789	.522	.842	.738	.799	.730	.849	.327
	Q (equipment)	.893	.826	.709	.795	.706	.790	.599	.705	.294
	R (hygiene)	.813	.782	.853	.792	.807	.836	.915	.852	.133

battery it primarily concerns items C (benefits/earnings stability), F (bullying/subordinates), M (contract), and N (security). What is significant for us, however, is that these are the items that underwent the most changes over the course of the years, including changes to the item concepts (C and F). It is therefore to be expected that these items would exhibit greater variability of factor loadings over time.

Confirmatory Factor Analysis (CFA)

We then conducted confirmatory factor analysis separately for each dataset using the maximum likelihood method. The test model corresponded to the instrument's original construct.[8] After giving consideration to the parameters of the

[8] Following from the results of the EFA, in the *importance* battery only one covariance was added between items O (chances) and K (development), since we can accept that opportu-

Table 8. CFA models: 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

Note: Model with covariance: O (chances) with K (development).

analysis, most notably the complexity of the model, the size of the datasets, and the number of years in the comparison, and also to the properties of the individual model fit statistics, the majority of which are sensitive to at least one of these parameters, we decided to use the following model fit statistics to assess the quality of the models: chi-square divided by the degrees of freedom ($\chi 2/df$), 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).[9] Chi-square statistics divided by the degrees of freedom should roughly have a value of 3. We will 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 at most 0.08 (it is sensitive to the complexity of the model), and the SRMR is at most 0.08 (ideally max. 0.06) [Hu and Bentler 1999; van de Schoot et al. 2012; Byrne 2010].

In the case of the *importance* battery, the tested model always represents the real data well: 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.

nities for personal development and further education are connected to a person's chances of finding another job in the labour market. As was then confirmed, adding it improved the results of the importance models in every year, as well as the results of the MG CFA.

[9] Here we also present the value of the Chi-square test (CMIN), which is sensitive to sample size. It tends to consider small differences on larger datasets to be significant (it is statistically significant), while on small samples it does not reject an inappropriate model [Kline 2010]. The primary reason we present CMIN is that most of the other indicators are derived from it.

Table 9. CFA models: Evaluation

Year	CMIN	df	p	CMIN /df	CFI	RMSEA	RMSEA 90% CI	SRMR	sample size
2009	436	120	.000	3.633	.941	.071	.064 – .079	.050	518
2011	484	120	.000	4.033	.898	.093	.084 – .101	.059	353
2013	284	120	.000	2.367	.948	.061	.052 – .071	.047	361
2014	880	120	.000	7.333	.958	.063	.059 – .067	.036	1611
2016	401	120	.000	3.342	.955	.062	.055 – .069	.037	611
2017	314	120	.000	2.617	.967	.053	.046 – .061	.038	568
2018	759	120	.000	6.325	.939	.079	.074 – .084	.051	855
2019	565	120	.000	4.708	.911	.100	.092 – .109	.054	368

The *evaluation* battery shows very similar results: the CFI is greater than 0.9 (except in 2011, when it is 0.002 lower), and the RMSEA always hovers around a value below 0.08, except in 2011 and 2019. However, the SRMR is below the threshold value of 0.08 in every year and in no year does it even rise above the value of 0.59. These results again point to possible minor problems in data comparability (e.g. for the year 2011 in the *evaluation* battery), but overall the model represents the data well in every year.[10]

Measurement invariance – MG CFA

The MG CFA models used to test measurement invariance were 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; Anýžová 2015; Svetina et al. 2019]). 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.[11]

[10] Based on the results we also tested a model with the added covariance of the items with the strongest relation. However, this did not lead to any fundamental improvement in the values of the model fit statistics; we therefore continued to work with the basic model.
[11] 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].

Table 10. MG CFA models over time: 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

Note: models with covariance: O (chances) with K (development).

Table 11. MG CFA models over time: Evaluation

Model	CMIN	df	P	CMIN /df	CFI	RMSEA	RMSEA 90% CI	AIC	BIC	SRMR
Configural	4121	960	.000	4.293	.945	.071	(.069–.073)	803468	807092	.044
Metric	4513	1044	.000	4.323	.940 (.001)	.071 (0)	(.069–.073)	803692 (224)	806765 (–327)	.057
Scalar	6612	1128	.000	5.862	.905 (.035)	.086 (.015)	(.084–.088)	805623 (–69)	808144 (1379)	.104

We will first examine the *importance* battery. The baseline configural model that tests whether the model represents the data well in all the compared years together 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 model, where there is a requirement that the factor loadings be equal, 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 years well and metric invariance is also confirmed.

We see a different situation in the scalar model, where the requirement is that the factor loadings be equal but also that all the item intercepts are equal as well. 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. What is important to emphasise here, however, is that the items that figure most often among the ones with the highest MI are C (benefits/earnings stability), F (bullying/subordinates), N (security), and O (chances). This is no surprise, as these are the items to which the biggest or the most changes were made over the years and the ones that the variability of the factor loadings had already drawn our attention to in the EFA.

After 2009 a change was made in the *evaluation* battery in which the concept was switched from *satisfaction* to *evaluation,* which could in theory have an adverse effect on even the configural invariance itself. Nevertheless, even in this battery configural invariance is achieved across all the observed years and the model fit statistics actually produce better results than in the *importance* battery. The CFI is above 0.94, the RMSEA is less than 0.08, and the SRMR is even less than 0.05. The metric model is also acceptable, as its results remain almost unchanged (the CFI shows a decrease of a mere 0.001, the RMSEA is unchanged, and the SRMR has an increase of 0.013). We can therefore declare the data in the *evaluation* battery to be metrically invariant as well.

Slightly worse results are observed again in the scalar model: although the CFI is below 0.09, the change to it is greater than 0.01 (0.03) and the RMSEA and the SRMR are more than 0.08. The model of full scalar invariance must be rejected. Here the highest MI are found for items C (benefits/earnings stability), F (bullying/subordinates), N (security), and L (independence). Again, these are the items that were already flagged as potentially problematic in previous analyses.

In order to proceed further by testing partial scalar invariance we would have to gradually release, step by step and across all the groups (years), the restrictions of parameter correspondence for the items with the highest modification indices. Given that we have a large number of items with a high MI, this would be a very lengthy process without any clear guidelines [Muthén and Asparouhov 2014]. The results do not state in detail which parameters should be relaxed first, what order to proceed in is up to the researcher to decide, which means that this is a rather subjective process that need not necessarily find support in the theory of the instrument. To arrive at an acceptable model, it would moreover be necessary to introduce numerous modifications, which could result in an excessively complicated and hard to interpret model. The final argument against taking this path is that it would be necessary to proceed differently for each of the batteries (each one has its own specific problems) and the outcome would thus be two different models; and this again might be inconsistent with the key theoretical concept of the instrument.

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

Approximate measurement invariance – the 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 metric or scalar invariance to be achieved – it is based on a 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 that are the most invariant and that are the most non-invariant over time. In order to obtain trustworthy alignment results and present meaningful comparisons between groups (years), the non-invariant rate must not exceed 25%.

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, which is well within the 25% cut-point. This is a very good result and it indicates that it is possible to compare the latent means derived from the alignment results between years [Muthén and Asparouhov 2014]. The most non-invariant items are F

Table 12. Alignment Method: Importance

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

* Cases with non-invariant parameters are highlighted.

(bullying/subordinates), I (harmonisation), and N (security), each of which was already identified as problematic in some of the previous methods.

Out of the 288 parameters for the *evaluation* battery, 22 intercepts and 9 factor loadings are non-invariant, which means that a total of 13.5% of the parameters are non-invariant. This is a slightly worse result than that of the *importance* battery, but it is still well below 25% and is therefore perfectly acceptable. Among those with the highest number of non-invariant parameters are items C (benefits/earnings stability), F (bullying/subordinates), I (harmonisation), and L (independence).

Because both batteries had good outcomes for all the years, according to the AM it is possible to compare the results (the factor means) of both batteries, and thus also to compare the overall result of the SQWLi indicator across all the observed years.

Table 13. Alignment method: Evaluation

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

* Cases with non-invariant parameters are highlighted.

Discussion

The analyses showed that SQWLi achieved configural and metric invariance over time, but not full scalar invariance, which is so essential for a meaningful comparison of single items means. A subject for further investigation could be to test for partial scalar invariance, now with more insight into problematic items whose parameters most likely need to be released.

On the other hand, the results of the approximate invariance approach used in this study show the SQWLi is at least approximately invariant over time. Our findings of non-invariance for both batteries were well within the recommended cut-point of 25%, which means factor means and variances estimated by the Alignment Method are plausible and can be compared between groups or years

Table 14. Measurement invariance testing: overview

	Problem expectation	Importance			Evaluation		
		EFA	MI	AM	EFA	MI	AM
A (earnings)	.	.	.	.	.	.	.
B (fair reward)	.	.	.	.	.	.	.
C (ben./e. stab.)	X	X	X	.	X	X	X
D (co-workers)	.	.	.	.	.	.	.
E (superiors)	X	.	.	.	.	.	.
F (bull./sub.)	X	X	X	X	X	X	X
G (time demands)	.	.	.	.	.	.	.
H (time flex.)	.	.	.	.	.	.	.
I (harmonization)	.			X	.	.	X
J (interestingness)	.	.	.	.	.	.	.
K (development)	.	.	.	.	.	.	.
L (independence)	X	.	.	.	.	X	X
M (contract)	X	X	.	.	X	.	.
N (security)	X	.	X	X	X	X	.
O (chances)	X	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.

in this case [Muthén and Asparouhov 2014]. These results indicate that if factor means can be compared, it is also possible to compare indicators for both batteries and the overall SQWLi indicator across all the observed years.

Evidently the most problematic in terms of invariance over time are the items that underwent a change in concept, which finding confirms our expectation. In the case of item F, in 2014 the concept was changed from *violence and bullying in the workplace* to *overall relations in the workplace*, and in 2019 it was changed again to *behaviour of subordinates towards superiors*. The concept of item C, *non-monetary benefits*, was in 2018 replaced with *earnings stability*. Also, a number of other changes were made to this item over the course of the years, when the term used to refer to benefits changed (from *benefits* to *bonuses*) and the examples and the overall wording of the item were also changed.

Although the different versions of the items always work well enough in the overall construct of the SQWLi, after a change in concept they in themselves are then measuring something different. Non-invariance was expected here; and demonstrating its occurrence specifically in the years in which these changes were made is in fact strong evidence to support the validity of the results of our analyses. Even without an analysis of invariance it is clear that it makes no sense to observe these items separately over time as a single phenomenon; consequently, when we confirm problems with the items in the invariance models, this does not call the instrument as a whole into question. On the contrary, 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.

The non-invariance of the other items can be explained by a change in their formulation, and again the year that our analysis identifies as a critical or problem year corresponds to the year in which a change was actually made in the questionnaire.

I (harmonisation): In 2017 the item's formulation was changed from *so that your work allows you enough time for your family, interests, and relaxation* and became *so that your work does not interfere with your personal time, i.e. time for family, interests or relaxation*.

L (independence): In 2014 the item's formulation was changed from *perform work independently* and became *to have the opportunity to decide on your own work tasks, organise your work independently*.

M (contract): In 2014 *type of economic activity* was a new aspect added to the characteristics of a person's employment relationship (*employee/self-employed*), while the aspect of whether the work is a *primary or secondary job* was removed; and in 2017 the aspect of *full- or part-time contract* was added.

N (security): In 2014 the item's formulation *job security* was changed to *employment security*, and in 2016 it was changed again and became *to be sure you don't lose your job*.

O (chances): The formulation of this item was changed frequently: *…opportunity and possibilities for employment in the labour market* (2009)*; …you gained an opportunity and possibilities for further employment in the labour market* (2011)*; … at work you were able to advance your chances…* (2014)*; …you sense from your work that you have a chance of finding further employment in the labour market* (2016); *…so that your work gives you the chance of further possible employment in the labour market* (2017).

Q (equipment): In 2014 the item's formulation was changed from *technical equipment in the workplace* to *technical equipment for work.*

All the items that we expected to have problems were indeed revealed to have problems by at least one of the methods used, except for item E (superiors). Most of them were directly uncovered by AM, some were only uncovered when we

used EFA or MG CFA. The analysis also showed that the items that were not expected to have problems because they did not undergo very pronounced or frequent changes over the years were not revealed to have problems in the invariance analysis; an exception to this is item Q (equipment). In return, it can also be noted that those items that were not found to be problematic by any of the methods of analysis were the very ones to which the least changes had been made over the course of the years.

That the main problems uncovered in the analysis can be explained by actual changes made to the instrument, and vice versa, is in our view a good result. We acknowledge, however, that we are unable to explain some questions. An example is item E (superiors), where we expected that the change in the item formulation from *behaviour of superiors to subordinates* to *behaviour of people in a higher-ranking position (superiors, clients, etc.) to those in a lower-ranking position* in 2014 and then the return to the original formulation in 2019 would lead to greater invariance problems. By contrast, we did not expect to have a problem with item Q (equipment), which throughout the observed period had its formulation changed only once, in 2014, from *technical equipment in the workplace* to *technical equipment for work*. And with item M (contract), for example, the problems that were revealed do not match up with the years in which the changes were made to the item as well as they do in the case of other items.

One reason for these unexpected results may be that the magnitude of the changes and their potential effect on invariance are subjectively determined by us. In fact, we make attempts to gauge the cognitive processes of respondents when they are answering a question – i.e. the interpretations that they make individually within the context of the overall list of questions and the many interactions of meanings that then occur between individual questions. In some cases, we have no clear guideline on how to determine whether, for instance, a particular change in formulation is insignificant or if it will cause a more significant change in the item's meaning; or whether a significant change in meaning will then perhaps alter the actual concept the item is measuring. It is possible that in one context even with a bigger linguistic change there will be no change in what an item means to respondents (E), while in a different context to change even a single word could significantly alter the item's meaning (Q). What follows from this is that we cannot in advance estimate absolutely all of the effects on invariance that may occur solely on the basis of a subjective assessment of the number and magnitude of the changes.

Because the overall results of the AM are good for both batteries, it is possible to compare the factor means and the indices for both batteries. With respect to individual items, however, it is necessary to proceed very cautiously. Given that neither full nor partial scalar invariance was confirmed by the MG CFA, it is not possible to compare the averages of all the items across all years. At the same time, the results of the AM provide information on specific items. In the *importance* battery there is full invariance between items B, G, J, K, L, P, and R,

there is minimal non-invariance for items A, M, O, C, D, H, and Q, items I and F are borderline non-invariant, and it is definitely not possible to compare item N over time. We have the same information also for the *evaluation* dimension, where items A, B, E, K, N, and R are fully invariant, there is a minimal share of non-invariance for items D, G, J, M, O, P, H, Q, there is borderline invariance for item C, while items F, I, and L are non-invariant over time.

Conclusion

The aim of this study was to test the quality of the SQWLi instrument in terms of its measurement invariance over time and through our analyses contribute to the discussion of measurement invariance as such. Our objective was to determine to what extent certain changes in a research instrument have an adverse effect on measurement invariance and thus on the comparability of time series data, and how strong an effect the different changes and modifications (that in reality for various reasons occur in the instrument over time) have on measurement invariance.

It is good to point out that an inverse reading of the results can answer the question of what kinds of changes and what instrument parameters, and under what conditions, do not lead to the incomparability of the measurement. Undoubtedly it is optimal to examine such questions under the specific conditions of a controlled experiment within the framework of an experimental and a control group schema. We have also conducted this type of examination of the SQWLi instrument in other studies, using more experimentally designed data (different versions of the instrument in single split sample surveys). This type of examination is clearly much better for assessing whether the measurement invariance is disrupted by the change in the instrument or by the change of perception of the concept(s) over time. In this study we sought to answer a broader research question and, most importantly, to do so under the more common conditions of real (non-experimental) surveys, where instruments often change in time for various reasons, and where there is need to assess the invariance of their results as well.

In the case of the SQWLi research instrument we were able to confirm that there was no negative effect on the configural invariance even when the concept of the battery as a whole was changed from *satisfaction* to *evaluation*, or when individual changes were made to the question wording or when the order of questions in the questionnaire's administration was altered. The metric invariance was not fatally affected even by the relatively pronounced changes made to the rating scales, where the range, polarity, and even the value labels were changed. An important finding here, however, is that in some cases the metric invariance can be increased ex post by means of the proper harmonisation of data. The risk to the invariance of the SQWLi data only begins to appear on the level of scalar invariance, and how serious the problems that arise are corresponds mostly to the size and number of changes that the individual items underwent

over the years. Based on the results of the AM, however, we can also say that even imperfections among individual items do not generate fatal problems for the instrument as a whole and that even in this case the overall invariance of the instrument is confirmed.

It is possible then to conclude that the SQWLi questionnaire is suitable for creating time series of an index of overall working-life quality or for creating individual indices of more general domains and of specific aspects (the ones that did not undergo concept changes and with a minimal share of non-invariance) of working life in the Czech Republic. It is possible to use data obtained by versions of the instrument since 2009 for this purpose, because the methodological and technical modifications that were made to it over the course of the years did not have a major effect on its measurement invariance and thus on the comparability of the results.

Jiří Vinopal *studied sociology at the Faculty of Arts, Charles University in Prague. He has been a researcher at the Institute of Sociology's Public Opinion Research Centre since 2001 and head of the Department of Sociology at the Faculty of Arts, Charles University, since 2013. His primary research focus is on the methodology and cognitive aspects of questionnaire surveys, public opinion research, and the methodology of researching working-life quality. He conducts projects of basic and applied research, publishes, and teaches courses on these topics. He is Deputy Chair of the Czech Sociological Association and is a member of Expert Evaluative Committee for the ETA Programme of the Technology Agency of the Czech Republic and a member of the Czech Editorial Board of* Sociologický časopis/Czech Sociological Review.

Kristýna Pospíšilová *has been a researcher at the Institute of Sociology of the Czech Academy of Sciences since 2015, and she also works as an expert consultant for the Czech Ministry of Labour and Social Affairs. She is currently completing her doctorate in sociology at Charles University in Prague with a specialisation in measurement equivalence over time. Her primary research interests are social inequalities in the labour market, childlessness and one-child families in the Czech Republic, and quantitative methodology, and she publishes on these topics as well.*

References

Anýžová, P. 2015. *Srovnatelnost postojových škál v komparativním výzkumu.* (The compatibility of attitude scales in comparative research) 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, https://doi.org/10.1080/10705511.2014.919210.

Byrne, B. M. 2008. 'Testing for Multigroup Equivalence of a Measuring Instrument: A Walk through the Process.' *Psicothema* 20 (4): 872–882.
Byrne, B. M. 2010. *Multivariate Applications Series. Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming.* 2nd ed. New York: Routledge.
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.
Cranny, C. J., P. C. Smith and E. F. Stone. 1992. *Job Satisfaction: How People Feel about Their Jobs and How It Affects Their Performance*. New York: Lexington Books.
Danna, K. and R. W. Griffin. 1999. 'Health and Well-Being in the Workplace. A Review and Synthesis of the Literature.' *Journal of Management* 25 (3): 357–384, https://doi.org/10.1177/014920639902500305.
Davidov, E. 2010. 'Testing for Comparability of Human Values across Countries and Time with the Third Round of the European Social Survey.' *International Journal of Comparative Sociology* 51 (3): 171–191, https://doi.org/10.1177/0020715210363534.
de Bustillo, R. M., E. F. Macias, J.-I. Antón and F. Esteve. 2011. 'Pluribus Unum? A Critical Survey of Job Quality Indicators.' *Socio-Economic Review* 9 (3): 447–475, https://doi.org/10.1093/ser/mwr005.
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.
Fuchs, T. 2009. 'Der DGB – Index Gute Arbeit.' Pp. 186–222 in *Arbeitsgestaltung als Zukunftsaufgabe Die Qualität der Arbeit*, edited by E. Kistler and F. Mußmann. Hamburg: VSA-Verlag.
Ghai, D. 2008. 'Decent Work: Concept and Indicators.' *International Labour Review* 142 (2): 113–145, https://doi.org/10.1111/j.1564-913X.2003.tb00256.x.
Hirschfeld, G. and R. von Brachel. 2014. 'Multiple-Group Confirmatory Factor Analysis in R – A Tutorial in Measurement Invariance with Continuous and Ordinal Indicators.' *Practical Assessment, Research and Evaluation* 19 (7). Retrieved 14 January 2020 (https://www.researchgate.net/profile/Gerrit_Hirschfeld/publication/266500478_Multiple-Group_confirmatory_factor_analysis_in_R_-_A_tutorial_in_measurement_invariance_with_continuous_and_ordinal_indicators/links/5433be3c0cf294006f71b7de/Multiple-Group-confirmatory-factor-analysis-in-R-A-tutorial-in-measurement-invariance-with-continuous-and-ordinal-indicators.pdf).
Hoppock, R. 1935. *Job Satisfaction*. New York and London: Harper.
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.
Judge, T. A., C. L. Hulin and R. S. Dalal. 2012. 'Job Satisfaction and Job Affect.' Pp. 496–525 in *Oxford Library of Psychology. The Oxford Handbook of Organizational Psychology, Vol. 1*, edited by S. W. J. Kozlowski. New York: Oxford University Press, https://doi.org/10.1093/oxfordhb/9780199928309.013.0015.
Kline, R. B. 2010. *Principles and Practice of Structural Equation Modeling*. 3rd ed. New York: Guilford Press.
Leschke, J., A. Watt and M. Finn. 2008. *Putting a Number on Job Quality? Constructing a European Job Quality Index*, ETUI-REHS, WP 2008.03. Retrieved 14 January 2020 (https://library.fes.de/pdf-files/gurn/00367.pdf).
Lomazzi, V. 2018. 'Measurement Invariance of Gender Role Attitudes in 59 Countries.' *Methods, Data, Analyses* 12 (1): 77–104.

Lowe, G. 2007. *21st Century Job Quality: Achieving What Canadians Want*. Ottawa: Canadian Policy Research Network.
Meredith, W. 1993. 'Measurement Invariance, Factor Analysis and Factorial Invariance.' *Psychometrika* 58: 525–543, https://doi.org/10.1007/BF02294825.
Muthén, B. and 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. and 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. and T. Asparouhov. 2014. 'IRT studies of Many Groups: The Alignment Method.' *Frontiers in Psychology* 5: 978, https://doi.org/10.3389/fpsyg.2014.00978.
Saris, W. E. and I. N. Gallhofer. 2014. *Design, Evaluation and Analysis of Questionnaires for Survey Research.* 2nd ed. Hoboken: Wiley, https://doi.org/10.1002/9781118634646.
Šeflová, K. 2016. 'Ekvivalence měření nástroje SQWLI v čase.' (Equivalence of SQWLI measurement over time) Dissertation, Faculty of Arts, Charles University. Retrieved 14 January 2020 (http://hdl.handle.net/20.500.11956/11700).
Sirgy, J. M., D. Efraty, P. Siegel and D.-J. Lee. 2001. 'A New Measure of Quality of Work Life (QWL) Based on Need Satisfaction and Spillover Theories.' *Social Indicator Research* 55 (3): 241–302, https://doi.org/10.1023/A:1010986923468.
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.
Svetina, D., L. Rutkowski and D. Rutkowski. 2019. 'Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using Mplus and the lavaan/sem Tools Packages.' *Structural Equation Modeling: A Multidisciplinary Journal* 0: 1–20, https://doi.org/10.1080/10705511.2019.1602776.
Swamy, D. R., T. S. Nanjundeswaraswamy and S. Rashmi. 2015 'Quality of Work Life: Scale Development and Validation.' *International Journal of Caring Sciences* 8 (2): 280–300.
Tangian, A. S. 2005. 'A Composite Indicator of Working Conditions in the EU-15 for Policy Monitoring and Analytical Purposes.' *WSI Working Papers 135*, The Institute of Economic and Social Research (WSI), Hans-Böckler-Foundation. Retrieved 14 January 2020 (https://ideas.repec.org/p/zbw/wsidps/135.html).
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.
van de Schoot, R., P. Lugtig and 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.
Vinopal, J. 2009. 'Erfahrungen mit der Messung der Qualität des Arbeitslebens in tschechischen Untersuchungen.' Pp. 143–177 in *Arbeitsgestaltung als Zukunftsaufgabe Die Qualität der Arbeit*, edited by E. Kistler and F. Mußmann. Hamburg: VSA-Verlag, https://doi.org/10.13060/00380288.2011.47.5.03.
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.' (Aspects of measuring the 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 and 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: coding and wording of items in the SQWLi questionnaire, 2019 version

Importance: '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.'

Evaluation: '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 and +5 VERY GOOD.'

Item code	Item wording
A (earnings)	a) The amount of earnings, i.e. the amount of your salary or wages.
B (fair reward)	b) …your work results are financially rewarded in a fair way.
C (benefits/earnings stability)	c) Earnings stability, …your salary is regular and stable.
D (co-workers)	d) Relationships between co-workers.
E (superiors)	e) Behaviour of subordinates towards superiors.
F (bullying/subordinates)	e) Behaviour of superiors towards subordinates.
G (time demands)	g) Total duration of working hours.
H (time flexibility)	h) The distribution of working hours during the day or week.
I (harmonization)	i) …your work does not interfere with your personal time, i.e. time for family, interests or relax.
J (interestingness)	j) …interesting work.
K (development)	k) …opportunities for further education and personal development at work.
L (independence)	l) …opportunity to decide on your own work tasks, organize your work independently.
M (contract)	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.
N (security)	n) …sure you don't lose your job.
O (chances)	o) …your work gives you the chance of further possible employment in the labour market.
P (h&s)	p) The level of occupational health and safety.
Q (equipment)	q) Technical equipment for work.
R (hygiene)	r) Cleanliness, tidiness and hygiene at work.